(F)utility Exposed

Abstract

In recent years, several authors have called to ground descriptive and normative decision theory on neuro-psychological measures of utility. In this paper, I combine insights from the best available neuro-psychological findings, leading philosophical conceptions of welfare and contemporary decision theory to rebut these prominent calls. I argue for two claims of general interest to philosophers, choice modellers and policy makers. First, severe conceptual, epistemic and evidential problems plague ongoing attempts to develop accurate and reliable neuro-psychological measures of utility. And second, even if these problems are solved, neuro-psychological measures of utility lack the potential to inform welfare analyses and policy evaluations.

Keywords: Utility; Choice; Welfare; Neuro-Psychology; Policy Evaluation.

Word Count: 4970

1. Introduction

Over the last two decades, neuro-psychological findings have enabled choice modellers to make significant advances in predicting and explaining individuals' choices (e.g. Camerer, 2008). Building on these advances, several authors have called to ground decision theory on neuro-psychological measures of utility (e.g. Dolan and Kahneman, 2008, Glimcher, 2010). The idea is that while standard decision theory takes rational agents to behave *as if* they maximize expected utility, recent neuro-psychological findings demonstrate that utility is *literally* computed by individuals' brains and can be *measured* in psycho-physical terms (e.g. Kable and Glimcher, 2009, Kahneman et al., 1997). These findings, in turn, are claimed to prompt revolutionary changes in both descriptive and normative decision theory. Descriptively, they allegedly provide "a neurobiological foundation for standard [decision] theory and a tool for measuring preferences neurobiologically" (Levy and Glimcher, 2012, 1027). Normatively, they purportedly undermine decision theory's "tight connection between choice and welfare" and enable researchers to identify "*true utility*, which encapsulates well-being" (Bernheim and Rangel, 2009, 51).

Calls to ground decision theory on neuro-psychological measures of utility have gained increasing popularity among choice modellers (e.g. Dolan and White, 2007), and support dissimilar evaluations of policies' welfare implications than standard decision theory (e.g. Angner, 2011, Hausman, 2010). In this paper, I combine insights from the best available neuro-psychological findings, leading philosophical conceptions of welfare and contemporary decision theory to rebut such prominent calls. I argue for two claims of general interest to philosophers, choice modellers and policy makers. First, severe conceptual, epistemic and evidential problems plague ongoing attempts to develop accurate and reliable neuro-psychological measures of utility. And second, even if these problems are solved, neuro-psychological measures of utility lack the potential to inform welfare analyses and policy evaluations. My point is not merely that different authors speak of 'true utility' in dissimilar senses or that current neuro-psychological measures are insufficiently developed. Rather, my thesis is that the search for true utility rests on ill-founded empirical and methodological presuppositions and that researchers should abandon this search altogether.[1]

The paper is organized as follows. Section 2 contrasts the notion of *decision utility* that figures prominently in standard decision theory and the notions of *true utility* put forward in the recent neuro-psychological literature. Section 3 examines the issue whether the available neuro-psychological findings enable researchers to develop accurate and reliable neuro-psychological *measures* of utility. Section 4 critically assesses the potential of neuro-psychological measures of utility to inform *welfare analyses* and *policy evaluations*. In doing so, it identifies and rebuts various objections that the proponents of true utility may put forward to vindicate their reliance on neuro-psychological measures.

[1] I speak of 'neuro-psychology' broadly to encompass several studies targeting individuals' hedonic experiences and the activation patterns of specific neural areas. Also, I use the expression 'inform welfare analyses and policy evaluations' to indicate the thesis that neuro-psychological measures of utility have direct descriptive and normative relevance for welfare analyses and policy evaluations.

2. Decision Utility versus True Utility

The notion of *decision utility* (henceforth, DU) refers to a mathematical representation of agents' preferences over choice options to be inferred from agents' choices between such options (e.g. Broome, 1991). DU figures prominently in the representation theorems at the core of standard decision theory (e.g. Savage, 1954). The idea is that if an agent's preferences satisfy specific axiomatic requirements (e.g. completeness, transitivity, independence), then the agent necessarily behaves *as if* she maximizes expected utility, i.e. of any two options, the one preferred is assigned higher expected utility. In this context, DU is regarded as derivative from choices and is ascribed no sensory or psychological properties (e.g. XXX, Okasha, 2016). That is to say, "when [the relevant] axioms are obeyed […] utility is not 'a feeling', or 'happiness', or 'a hedonic impulse' [but is just a mathematical index of] 'choice'" (Glimcher, 2009, 505).

The notion of *true utility* (henceforth, TU) refers to a measurable psycho-physical magnitude rather than a mathematical construct such as DU. In the recent neuro-psychological literature, various notions of TU have been proposed. The notions of *experienced utility* and *neural utility* are especially influential in this context. These two notions relate utility to empirical facts about individuals' neuro-psychology rather than to a mathematical index of individuals' choices. More specifically, the notion of *experienced utility* (henceforth, EU) refers to a psycho-physical magnitude that putatively reflects individuals' hedonic experiences and causally influences individuals' choices (e.g. Kahneman, 2000). EU figures prominently both in psychological studies of individuals' hedonic experiences and in some philosophers' interpretations of decision theory (e.g. Dietrich and List, 2016). In fact, EU is often claimed to resemble early utilitarian philosophers' notion of utility (e.g. Kahneman et al., 1997, on Bentham). The notion of *neural utility* (henceforth, NU) refers to the activation patterns of specific neural areas and the rewards' subjective values allegedly computed and integrated by such areas (e.g. Glimcher, 2010). NU figures prominently in the specialized neuro-psychological and neuroeconomic literature. In this literature, there is growing consensus that activations in the ventromedial prefrontal cortex/orbitofrontal cortex (henceforth, vmPFC/OFC) track rewards' subjective values across multiple reward types (e.g. Levy and Glimcher, 2012). The idea is that these subjective values "are exactly the kind of value representation posited by most decision theories […] analogous to what economists call 'utilities'" (Kable and Glimcher, 2009, 734).

3. True Utility: Descriptive Evaluation

In this section, I examine the issue whether the available neuro-psychological findings enable researchers to develop accurate and reliable neuro-psychological measures of utility. I argue that ongoing attempts to develop such measures face at least four major challenges: the *challenge from measurement divergences*, the *challenge from indeterminate referents*, the *challenge from unclear unit of measurement*, and the *challenge from conceptual barriers*. Some of these challenges (e.g. measurement divergences) affect also DU measures. Other challenges, instead, specifically target TU (rather than DU) measures, and license selective scepticism regarding TU measures.

3.1 The *challenge from measurement divergences* builds on the profound divergences between measurements that purportedly target the same notions of TU to question leading authors' aim to identify "a single, unifying concept that motivates all human choices and registers all relevant feelings and experiences" (Kahneman and Krueger, 2006, 4). This challenge targets both EU and NU measurements.

EU measurements can vary dramatically depending on what methods one uses to measure EU (e.g. XXX on divergences between physiological indices and individuals' reports of their own hedonic experiences). Moreover, significant divergences arise for the same methods depending on whether one focuses on *instant utility* measures, which target momentary evaluations of ongoing hedonic experiences, *remembered utility* measures, which target retrospective evaluations of past hedonic experiences, or *anticipated utility* measures, which target predictions concerning future hedonic experiences (e.g. Dolan and Kahneman, 2008, on individuals' failures to reliably predict their own future EU). These divergences arise not just from putative limitations of current measurement methods, but also from specific properties of the notions of TU supposedly targeted by neuro-psychological measures. For instance, the act of remembering past experiences typically involves active reconstruction (rather than passive recall) of these experiences. As a result, remembered utility measurements frequently fail to accurately reflect how pleasant past events were while being experienced (e.g. Kahneman, 2000, on so-called duration neglect). These concerns exacerbate when one considers that EU measurements can vary remarkably because of factors that are not directly related to the valence/phenomenal qualities of individuals' hedonic experiences (e.g. Fleurbaey, 2009, on temporary weather conditions).

Regarding NU, at least four features of the human neural architecture can generate significant divergences between different NU measurements: (1) rewards' subjective values depend not only on rewards' *objective attributes*, but also on interpersonally and intertemporally variable *reference points*, and different measures diverge as to how these reference points are affected by individuals' available, anticipated and recently obtained rewards (e.g. Glimcher, 2010); (2) to encode a potentially infinite range of reward values in a finite range of activations, the neural substrates of reward valuation do not represent rewards' *absolute* values, but rely on *normalization* processes that hinder the suitability of observed neural activations to serve as reliable proxies for rewards' subjective values across individuals and measurement contexts (e.g. Louie et al., 2013); (3) the neural substrates of reward valuation routinely *remove* subsets of incoming information streams (e.g. regularity-induced redundancies) to maximize the efficiency of information coding, thereby increasing neural activations' *independence* from the available rewards and stimuli (e.g. Padoa-Schioppa, 2011); (4) the brain does not *invariably* evaluate the impact of all factors that causally influence individuals' choices in terms of a direct computation of rewards' subjective values (e.g. Ross, 2008), so "there are many behaviors that can be described as a decision but are not [based on rewards'] subjective values" (Padoa-Schioppa and Schoenbaum, 2015, 17).

3.2 Challenge from indeterminate referents. The proponents of TU often leave it indeterminate whether TU tracks only individuals' psycho-physical *feelings* (*hedonic* interpretations of TU) or also their cognitive/evaluative *attitudes* regarding such feelings (*attitudinal* interpretations of TU). In fact, even the same authors frequently fail to specify which of these interpretations of TU they adopt (e.g. Angner, 2011, on Kahneman, 2000). This lack of specificity complicates attempts to provide accurate and reliable neuro-psychological measures of utility in at least three respects. First, neuro-psychological

measurements *significantly vary* depending on whether one adopts hedonic or attitudinal interpretations of TU (e.g. Crisp, 2006, on hedonic and attitudinal EU; Berridge and O'Doherty, 2014, on hedonic and attitudinal NU). Second, authors who agree on whether to adopt hedonic (or attitudinal) interpretations of TU frequently disagree regarding *what set* of feelings (or attitudes) are tracked by TU (e.g. Loewenstein and Ubel, 2008). And third, attempts to determine the referents of TU pose the following *dilemma* to the proponents of TU. On the one hand, hedonic interpretations of TU fail to track many of the referents that researchers associate with TU (e.g. Heathwood, 2006, on physiological indices' failure to accurately track the influence of individuals' desires on the valence/phenomenal qualities of their own hedonic experiences). On the other hand, attitudinal interpretations of TU face severe measurability concerns. To illustrate these concerns, consider measurements of attitudinal EU.

There are several reasons why measurements of attitudinal EU may fail to accurately track the valence/phenomenal qualities of individuals' hedonic experiences: (1) the valence/phenomenal qualities of some hedonic experiences are hard to *assess* for individuals (e.g. Haybron, 2005, on various moods), and individuals' *abilities* to assess such valence/phenomenal qualities vary across individuals and situations in ways that are difficult to monitor (e.g. Fleurbaey, 2009); (2) the *scales* individuals use to evaluate their own hedonic experiences vary across time and situations to a degree that hampers both the intrapersonal and the interpersonal comparability of individuals' evaluations (e.g. Loewenstein and Ubel, 2008, on individuals' tendency to normalize hedonic evaluations of their own health conditions based on implicit standards of comparison); (3) individuals' tendency to *adapt* to affects and circumstances can alter their hedonic evaluations in ways that elude both their own awareness and physiological indices of EU (e.g. Van der Deijl, 2017); (4) *social norms and pressures* can demonstrably modify individuals' hedonic evaluations in the absence of experimentally measurable variations in their hedonic experiences (e.g. Dolan and White, 2007); (5) individuals' hedonic evaluations can be significantly affected by their own *expectations* concerning how they ought (or are likely) to feel in specific circumstances (e.g. Haybron, 2007).

3.3 Challenge from unclear unit of measurement. The proponents of TU frequently maintain that neuro-psychological measures of utility enable researchers to "objectively compare mental state between individuals" (Glimcher, 2010, 425) and provide "policymakers [with] a standard unit of measurement for comparisons of well-being across domains" (Dolan and White, 2007, 71). However, it remains unclear what unit of measurement is supposed to ground these comparisons. Moreover, attempts to identify such unit of measurement face daunting difficulties in case of both EU and NU.

EU is often claimed to track a common quality of pleasantness or desirability allegedly shared by individuals' hedonic experiences (e.g. Kahneman and Krueger, 2006). The qualitative differences between distinct hedonic experiences do not exclude the possibility that these experiences vary in magnitude along some common scale (e.g. Kagan, 1992). Still, the proponents of EU have hitherto failed to specify *what* this putative common scale consists in. In fact, several authors doubt the prospects of identifying a common scale for measuring the pleasantness or desirability of individuals' hedonic experiences (e.g. Crisp, 2006). A proponent of EU may respond that even if researchers are unable to identify a common scale for measuring the pleasantness or desirability of individuals' hedonic experiences, EU can be taken to track all experiences toward which individuals have some positive attitude (e.g. Heathwood, 2006). However, to address the challenge from unclear unit of measurement, this response must be supplemented with a

clear specification of *which* positive attitudes are tracked by EU and on what scale these attitudes are to be *measured*. In the absence of this specification, lumping all experiences toward which individuals have some positive attitude under the label of EU obscures (rather than solves) the measurability problems posed by the diversity (and potential incommensurability) of individuals' hedonic experiences.

As to NU, several authors hold that to enable individuals to choose between complex (e.g. multi-attribute) alternatives, individuals' neural architecture must compute and compare rewards' subjective values on a common neural scale (e.g. Landreth and Bickle, 2008). In particular, leading researchers concur that anatomically delimited activations in vmPFC/OFC "encode the subjective values of different types of rewards on a neural common scale" (Levy and Glimcher, 2012, 1027). There are at least four reasons to doubt the proffered identifications of a common neural scale: (1) the neural substrates of reward valuation are *highly distributed*, with several areas besides vmPFC/OFC being involved in computing and comparing value-related signals across choice situations (e.g. Rushworth et al., 2012); (2) activations in vmPFC/OFC contribute to *multiple functions* (e.g. Hunt et al., 2012), which hinders attempts to establish that such activations encode rewards' subjective values rather than contribute to some other function (e.g. Weiskopf, 2016); (3) most of the proffered correspondences between individuals' choices and the neural substrates of reward valuation target *stimulus-bound* choices rather than the non-stimulus-bound choices individuals face in non-laboratory settings (e.g. Ross, 2011), and researchers lack convincing evidence that such correspondences generally hold also for non-stimulus-bound choices (e.g. XXX); (4) many correspondences between individuals' choices and the neural substrates of reward valuation target *rather coarse* characterizations of these substrates, and break down when more fine-grained characterizations of such substrates are targeted (e.g. Padoa-Schioppa and Schoenbaum, 2015, on the dissimilar functional roles of distinct sub-regions of OFC). As a result, such correspondences are more plausibly regarded as consequences of researchers' taxonomic practice rather than independent empirical evidence of a common neural scale.

3.4 Challenge from conceptual barriers. In the specialized philosophical literature, various conceptual differences between DU and EU have been identified (e.g. Okasha, 2016). Here I identify and explicate three major conceptual differences between DU and NU: (1) DU represents *ordinal* relations between choice options rather than unique *cardinally* comparable magnitudes. For this reason, a neuron cannot be plausibly taken to represent the utility of a choice option even if this neuron's firing rate is linearly proportional to the value of such option (e.g. Glimcher, 2009, 506); (2) while DU is a mathematical index of individuals' choices, neural measures of rewards' subjective values are empirically *dissociable* from choices and can *causally* influence choices (e.g. Ross, 2011); (3) there is ongoing debate as to *how many* kinds of value-related *signals* are computed in the human neural architecture and how exactly these signals are integrated intro overall measures of value (e.g. Padoa-Schioppa, 2011, on action-related values and values computed in the space of goods irrespective of sensorimotor contingencies; Padoa-Schioppa and Schoenbaum, 2015, on values learned and retrieved from memory and values computed at the time of choice). These conceptual differences do not prevent researchers from identifying statistically significant correlations between individuals' choices and specific areas' activation patterns (e.g. Kuorikoski and Marchionni, 2016). However, they cast serious doubt on leading authors' claim that rewards' subjective values "are exactly the kind of value representation posited by most decision theories […] analogous to what economists call 'utilities'" (Kable and Glimcher, 2009, 734).

4. True Utility: Normative Evaluation

In this section, I critically assess the potential of neuro-psychological measures of utility to inform welfare analyses and policy evaluations. In doing so, I identify and rebut various objections that the proponents of TU may put forward to vindicate their reliance on neuro-psychological measures. Standard decision theory commonly takes individuals to be well off to the extent that their actual, informed or ideal *preferences* are satisfied (e.g. Hausman and McPherson, 2009). An individual's preferences count as satisfied when the state of affairs targeted by these preferences obtains. Knowing that one's preferences are satisfied in this sense may give one a feeling of satisfaction. Yet, preference satisfaction does not have to involve any feeling of satisfaction, and may even obtain when one is unaware that her preferences are satisfied (e.g. Hausman, 2010). For their part, the proponents of TU regard welfare as measurable on interpersonally comparable scales and rely on neuro-psychological measures of utility as the relevant *indexes* of welfare (e.g. Kahneman and Krueger, 2006). The idea is that neuro-psychological measures of utility provide accurate and reliable indexes of welfare that enable policy makers to establish "when people […] choose what is best for them [and] what good policies follow" (Camerer, 2008, 416; Bernheim and Rangel, 2009, 85).[2]

I already discussed in Section 3 some of the limitations that plague current neuro-psychological measures of utility. Below I assume, for the sake of argument, that researchers are able to overcome these limitations. I then argue that even if such limitations are overcome, informing welfare analyses and policy evaluations requires researchers to solve at least five major problems: (1) provide plausible and informative criteria to identify what welfare-relevant operations (if any) are tracked by neuro-psychological measures of utility (*identification problem*); (2) make sure that no welfare-relevant operations are neglected by these measures (*exclusion problem*); (3) make sure that no welfare-irrelevant operations are tracked by such measures (*purification problem*); (4) determine what weights to ascribe to the welfare-relevant operations purportedly tracked by neuro-psychological measures (*determination problem*); (5) aggregate the weighted measures of welfare-relevant operations thus determined into summative welfare indexes for the targeted individuals/population subsets (*aggregation problem*).

Some of these problems affect also DU measures (e.g. Bernheim, 2009). However, the proponents of TU have hitherto failed to demonstrate that neuro-psychological measures solve such problems better than DU measures (e.g. XXX). Moreover, each of those problems poses formidable challenges to the proponents of TU. More specifically: (1) solving the identification problem requires researchers to make controversial *normative/axiological presuppositions* concerning the nature and the referents of welfare (e.g. is welfare related more closely to EU or NU? Is welfare tracked more reliably by momentary, retrospective or anticipated evaluations of individuals' experiences?); (2) solving the exclusion problem is complicated by the fact that even basic welfare-relevant operations (e.g. anticipating or reminiscing specific experiences) typically engage *several* neuro-psychological processes besides those targeted by neuro-psychological measures of utility (e.g. Muldoon and Bassett, 2016); (3) solving the purification problem is

[2] I use 'well-being' and 'welfare' interchangeably to indicate what makes a life good for the individual living such life (e.g. Griffin, 1986). Also, I speak of 'conceptions of welfare' to indicate theories that specify both which goods/experiences are welfare-enhancing and what properties make such goods/experiences welfare-enhancing (e.g. Kagan, 1992).

complicated by the fact that the neuro-psychological processes that purportedly contribute to welfare-relevant operations *also* contribute to several *non*-welfare-relevant operations (e.g. Kable and Levy, 2015); (4) solving the determination problem requires researchers to establish what *normative/axiological significance* should be ascribed to neuro-psychological measures, which "are explicitly positive in nature" (Glimcher, 2010, 412); (5) solving the aggregation problem is complicated by the wide *diversity* (and potential incommensurability) of welfare-relevant goods/experiences (e.g. Hausman, 2012) and by the limited *interpersonal* comparability of neuro-psychological measures (Section 3).

A proponent of TU may grant that problems (1)-(5) challenge attempts to provide *perfectly* accurate and reliable neuro-psychological indexes of welfare. However, she may object that neuro-psychological indexes of welfare are *sufficiently* accurate and reliable to evaluate most policies' welfare implications (e.g. Kahneman, 2000). There are at least three reasons to doubt that neuro-psychological indexes of welfare are sufficiently accurate and reliable to evaluate most policies' welfare implications. First, neuro-psychological indexes track a *relatively narrow subset* of the goods/experiences that on most conceptions of welfare count as welfare-enhancing (e.g. Alexandrova, 2017, on achievement, freedom and justice). Second, the goods/experiences that on most conceptions of welfare count as welfare-enhancing frequently contribute to welfare *irrespective* of what impact they happen to have on neuro-psychological indexes (e.g. Hausman and McPherson, 2009). And third, neuro-psychological indexes track several goods/experiences that on most conceptions of welfare do *not* count as welfare-enhancing (e.g. pleasures based on mistaken beliefs) and may even detract from welfare (e.g. pleasures derived from immoral activities).

A proponent of TU may object that neuro-psychological indexes provide informative insights regarding policies' welfare implications because they reliably track *necessary components* of welfare (e.g. Dolan and Kahneman, 2008, for statistically significant correlations between neuro-psychological indexes and various indicators of health). However, it is dubious that neuro-psychological indexes *reliably* track necessary components of welfare (e.g. dopaminergic activations targeted by such indexes can be induced by addictive substances that demonstrably hamper individuals' welfare). Moreover, some goods/experiences may be necessary components of welfare, yet fall short of providing *reliable* (or even approximate) indicators of welfare (e.g. think of oxygen availability). This, in turn, constrains neuro-psychological indexes' potential to inform the evaluation of policies' welfare implications even under the assumption that these indexes reliably track necessary components of welfare. For example, consider Loewenstein and Ubel's claim that "evaluations of welfare will inevitably have to be informed by a combination of [choice-based and neuro-psychological indexes] patched together in a fashion that depends on the specific context" (2008, 1795). This claim does not provide any informative insights regarding policies' welfare implications unless one supplements it with more detailed indications regarding what indexes are considered, how these indexes are 'patched together' and how 'the specific context' determines the normative/axiological significance of the goods/experiences tracked by such indexes.

A proponent of TU may object that neuro-psychological indexes enable researchers to evaluate policies' welfare implications by categorizing individuals into distinct behavioural/neuro-psychological *types* (e.g. Kable and Levy, 2015, for statistically significant correlations between individuals' risk propensities and morphometric indicators such as gray matter volume). Even so, it remains unclear how exactly these categorizations are supposed to inform researchers' evaluations of policies' welfare

implications. For showing that a policy is welfare-enhancing for specific behavioural/neuro-psychological types of individuals requires researchers to establish systematic correspondences between these behavioural/neuro-psychological types and the policy's welfare implications for individuals of those types. And a given policy may have dissimilar welfare implications for individuals of the same behavioural/neuro-psychological types (e.g. XXX). Therefore, researchers' hypothesized ability to categorize individuals into distinct behavioural/neuro-psychological types falls short of indicating that they can accurately evaluate policies' welfare implications for individuals of these types.

A proponent of TU may further object that neuro-psychological indexes enable researchers to *resolve disagreements* concerning policies' welfare implications *without* having to make normative/axiological assumptions (e.g. Bernheim and Rangel, 2009, 85, for the claim that neuro-psychological indexes can "in principle [resolve] any serious disagreement" as to whether specific policies enhance individuals' welfare). Still, justifiably inferring that policies which enhance the value of specific neuro-psychological indexes also enhance individuals' welfare requires researchers to specify how such indexes map on the notion of welfare. In fact, even this specification does not *per se* enable researchers to establish "what good policies follow" from neuro-psychological indexes (Camerer, 2008, 416). For demonstrating that a particular policy is justified requires one to show not just that this policy enhances individuals' welfare, but also that such policy does not involve morally unacceptable violations of normatively significant values such as autonomy or consent (e.g. Hausman, 2012). In this respect, it is telling that distinct proponents of TU disagree regarding what policies (if any) are supported by the neuro-psychological indexes they advocate (e.g. Kahneman and Sugden, 2005).

5. Conclusion

The critique put forward in this paper points to two major reasons to think that the search for the Holy Grail of 'true utility' should be abandoned. First, severe conceptual, epistemic and evidential problems plague ongoing attempts to develop accurate and reliable neuro-psychological measures of utility. And second, even if these problems are solved, neuro-psychological measures of utility lack the potential to inform welfare analyses and policy evaluations. These two reasons do not prevent researchers from identifying systematic correspondences between individuals' choices and psycho-physical magnitudes amenable to neuro-psychological measurements. Yet, if the present critique is correct, such correspondences provide neither 'a neurobiological foundation for standard [decision] theory' nor a notion of 'true utility [that] encapsulates well-being'.

REFERENCES

Alexandrova, A. 2017. Can the science of well-being be objective? *British Journal for the Philosophy of Science*. In press.

Angner, E. 2011. Are subjective measures of well-being 'direct'? *Australasian Journal of Philosophy*, 89, 115-130.

Bernheim, D. 2009. On the potential of neuroeconomics. *American Economic Journal*, 1, 1-41.

Bernheim, D. and Rangel, A. 2009. Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics. *Quarterly Journal of Economics*, 124, 51-104.

Berridge, K. and O'Doherty, J. 2014. From expected utility to decision utility. In *Neuroeconomics: Decision Making and the Brain*. 2nd Ed. 335-354. Elsevier.

Broome, J. 1991. Utility. *Economics and Philosophy*, 7, 1-12.

Camerer, C. 2008. Neuroeconomics. *Neuron*, 60, 416-419.

Crisp, R. 2006. Hedonism Reconsidered. *Philosophy and Phenomenological Research*, 73, 619-645.

Dietrich, F. and List, C. 2016. Mentalism versus behaviourism in economics: a philosophy-of-science perspective. *Economics and Philosophy*, 32, 249-281.

Dolan, P. and Kahneman, D. 2008. Interpretations of utility and their implications for the valuation of health. *Economic Journal*, 118, 215-234.

Dolan, P. and White, M. 2007. How can measures of subjective well-being be used to inform public policy? *Perspectives on Psychological Science*, 2, 71-85.

Fleurbaey, M. 2009. Beyond GDP: the quest for a measure of social welfare. *Journal of Economic Literature*, 47, 1029-1075.

Glimcher, P. 2009. Choice: towards a standard back-pocket model. In *Neuroeconomics: Decision Making and the Brain*, 503-521. Elsevier.

Glimcher, P. 2010. *Foundations of Neuroeconomic Analysis*. Oxford University Press.

Griffin, J. 1986. *Well-Being: Its Measure and Importance*. Clarendon Press.

Hausman, D. 2010. Hedonism and welfare economics. *Economics and Philosophy*, 26, 321-344.

Hausman, D. 2012. *Preference, Value, Choice, and Welfare*. Cambridge University Press.

Hausman, D. and McPherson, M. 2009. Preference satisfaction and welfare economics. *Economics and Philosophy*, 25, 1-25.

Haybron, D. 2005. On being happy or unhappy. *Philosophy and Phenomenological Research*, 71, 287-317.

Haybron, D. 2007. Do we know how happy we are? *Nous*, 41, 394-428.

Heathwood, C. 2006. Desire satisfactionism and hedonism. *Philosophical Studies*, 128, 539-563.

Hunt, L., Kolling, N., Soltani, A., Woolrich, M., Rushworth, M. and Behrens, T. 2012. Mechanisms underlying cortical activity during value-guided choice. *Nature Neuroscience*, 15, 470-476.

Kable, J. and Glimcher, P. 2009. The neurobiology of decision. *Neuron*, 63, 733-745.

Kable, J. and Levy, I. 2015. Neural markers of individual differences in decision-making. *Current Opinion in Behavioral Sciences*, 5, 100-107.

Kagan, S. 1992. The limits of well-being. *Social Philosophy and Policy*, 9, 169-189.

Kahneman, D. 2000. Experienced utility and objective happiness. In *Choices, Values, and Frames*, ch.37. Cambridge University Press.

Kahneman, D. and Krueger, A. 2006. Developments in the measurement of subjective wellbeing. *Journal of Economic Perspectives*, 20, 3-24.

Kahneman, D. and Sugden, R. 2005. Experienced utility as a standard of policy evaluation. *Environmental and Resource Economics*, 32, 161-181.

Kahneman, D., Wakker, P. and Sarin, R. 1997. Back to Bentham? Explorations of experienced utility. *Quarterly Journal of Economics*, 112, 375-406.

Kuorikoski, J. and Marchionni, C. 2016. Evidential diversity and the triangulation of phenomena. *Philosophy of Science*, 83, 227-247.

Landreth, A. and Bickle, J. 2008. Neuroeconomics, neurophysiology and the common currency hypothesis. *Economics and Philosophy*, 24, 419-429.

Levy, D. and Glimcher, P. 2012. The root of all value: a neural common currency for choice. *Current Opinion in Neurobiology*, 22, 1027-1038.

Loewenstein, G. and Ubel, P. 2008. Hedonic adaptation and the role of decision and experience utility in public policy. *Journal of Public Economics*, 92, 1795-1810.

Louie, K., Khaw, M. and Glimcher, P. 2013. Normalization is a general neural mechanism for context-dependent decision making. *PNAS*, 110, 6139-6144.

Muldoon, S. and Bassett, D. 2016. Network and multilayer network approaches to understanding human brain dynamics. *Philosophy of Science*, 83, 710-720.

Okasha, S. 2016. On the interpretation of decision theory. *Economics and Philosophy*, 32, 1-25.

Padoa-Schioppa, C. 2011. Neurobiology of economic choice: a good-based model. *Annual Review of Neuroscience*, 34, 333-359.

Padoa-Schioppa, C. and Schoenbaum, G. 2015. Dialogue on economic choice, learning theory, and neuronal representations. *Current Opinion in Behavioral Sciences*, 5, 16-23.

Ross, D. 2008. Two styles of neuroeconomics. *Economics and Philosophy*, 24, 473-483.

Ross, D. 2011. Estranged parents and a schizophrenic child: choice in economics, psychology and neuroeconomics. *Journal of Economic Methodology*, 18, 217-231.

Rushworth, M., Kolling, N., Sallet, J. and Mars, R. 2012. Valuation and decision-making in frontal cortex. *Current Opinion in Neurobiology*, 22, 946-955.

Savage, L. 1954. *The Foundations of Statistics*. John Wiley and Sons.

Van der Deijl, W. 2017. Which problem of adaptation? *Utilitas*, 29, 474-492.

Weiskopf, D. 2016. Integrative modeling and the role of neural constraints. *Philosophy of Science*, 83, 674-685.